Imperial College London

Imperial College London

Department of Life Science

Genomic Architecture of Parallel Ecological Divergence in Littorina saxatilis

Author:
Rui Zhang

CID:
01907894

Email:
rui.zhang20@imperial.ac.uk

Supervisor:
Matteo Fumagalli
m.fumagalli@imperial.ac.uk

Co-supervisor:
Francesca Raffini
f.raffini@sheffield.ac.uk

Roger K. Butlin
r.k.butlin@sheffield.ac.uk

Date: 8/26/2021

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Science at Imperial College London

Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

I declare that this thesis is solely composed by my own work. The data was provided by co-supervisor Francesca Raffini and Roger K. Butlin from The University of Sheffield. The preliminary data processing was conducted by co-supervisor Francesca Raffini. I only executed data analysis and results visualization. Supervisor Matteo Fumagalli provided a workshop on how to use ANGSD v.0.934. Before August 2021, this thesis has not been published in any journal.

Signature:

Date: 8/26/2021

Contents

Abstract

Parallel adaptation of *Littorina saxatilis* in heterogeneous environments has been researched by many scientists all around the world, since it is a good subject to explore sympatric speciation and natural selection. This species has evolved two ecotypes adapted to micro-habitats: crab ecotype and wave ecotype. The aim of this research is to find underlying mechanisms of parallel ecological divergence with gene flow and the evolvement of reproductive isolation with inter-breeding. Low coverage whole genome sequencing and related data analysis tools were used on 73 samples of snails and the research on hybrids is the emphasis.

In the Spanish system of *Littorina saxatilis*, snails are divided into two distinct genetic clusters instead of a cline in Sweden. The crab group and hybrid group are genetically closer and they compose the first cluster, a cline. The wave group forms the second homogeneous cluster. Also, there is an obvious directional introgression from wave into crab at a low rate which destroys high Linkage Disequilibrium (LD) in crab ecotype. Differentiation analysis indicates that most loci are polymorphic and not fixed suggesting that loci are under both divergent selection and balancing selection. Usually, chromosome inversion candidates show higher FST and LD compared with whole chromosome, while PBS and genetic diversity (pi) are more diverse. The crab group has the highest LD and slowest LD decay rate which hints at stronger selection and more polymorphic inversions. It is hard to find new chromosome inversions just from LD heat map, but some chromosomes demonstrate earlier points dispersal in LD decay suggesting short inversions. In conclusion, this research found some new chromosome inversions candidates to be verified and tried to explain the process of ecological divergence after the phase of a cline in terms of genetic structure and selection.

Keywords

Whole genome sequencing, parallel evolution, local adaptive, genomic architecture, gene flow, reproductive isolation

1 Introduction

How speciation occurs is a fundamental biological problem. There are three traditional geographic modes of speciation: allopatric speciation, parapatric speciation, and sympatric speciation (Fitzpatrick et al. 2008). Sympatric speciation is fascinating due to its difficulty in theory explanation and definition. Parallel ecological divergence is one of the most interesting subjects of sympatric speciation. A monomorphic population could split into two coexisting phenotypic clusters (ecotypes) under directional selection of ecological environments explained by the theory of adaptive dynamics (Dieckmann & Doebeli 1999). It is well established that divergent natural selection in heterogeneous ecological environments is likely to be an impetus for local adaptation and subsequent reproductive isolation even speciation (Schluter 2009). Repeated occurrence of ecological divergence in discrete populations in contrasting habitats gives an opportunity to explore speciation mechanisms. Moreover, how reproductive isolation evolves with interbreeding, which means the presence of gene flow impeding speciation, is more informative to understand initial steps of speciation than the study on already separated species (Kirkpatrick & Ravigné 2002).

Littorina saxatilis is a common rocky-shore gastropod with short life and high lifetime dispersal due to ovoviviparity (Reid 1996). This species has a strong adaptation ability. Even small patches of local habitat could promote a distinct ecotype under biotic or abiotic pressure (Johannesson 2016). It has evolved two ecotypes under the selection of wave exposure and crab predation in many locations such as northwest Spain, west coast of Sweden, and northeast coast of England (Butlin et al. 2008). Wave ecotype is smaller and has a thinner shell with a larger aperture, while crab ecotype is bigger and has a thicker shell with a smaller aperture (Johannesson et al. 2010). The morphology and behavior differences between two ecotypes are inheritable although with a minor phenotypic plasticity (Galindo et al. 2009). *Littorina saxatilis* has developed partially reproductive isolation in the face of hybridization (Grahame et al. 2006). This proposition is supported by that gene flow in contact zones was 10-30 % of gene flow within ecotypes and that the genetic relationship between crab ecotype and wave ecotype in the same location is closer than the genetic relationship between the same ecotype in different locations (Panova et al. 2006). Also, Beaumont (Beaumont 2010) got strong support for parallel and local divergence in this species instead of old allopatric divergence of ecotypes followed by secondary overlap and gene flow using approximate Bayesian computation (ABC) approach. Therefore, it is a preferable model system to study the underlying mechanisms, particularly genomic architecture variation, of parallel adaptive divergence with the existence of gene flow, which is sympatric or parapatric speciation instead of allopatric speciation.

There are many studies focusing on the history of *Littorina saxatilis* dispersal, colonization, formation of ecotypes, and evolution of reproductive barriers between two ecotypes under gene flow. This species is inferred to survive from the last glacial period in a northern latitudes refugia based on phylogeographic data (Panova et al. 2011). Later colonization seemed to be achieved by rafting of single females whose brood pouch carried hundreds of embryos sired by

even over 20 males. This feature allows a single female to be a founder group with diversified genetic variation and a new population is established rapidly once releasing embryos (Rafajlovic et al. 2013). Crab is regarded as the driving force of crab ecotype formation. After snails' colonization, the predator of snails arrived at the same location later due to their higher minimum temperature requirement compared with snails. In the ABC model established by Beaumont (Beaumont 2010), ecotypes separation is a relatively recent incident. The ecotype separation time is estimated at only around 10 % of local population age. Furthermore, the formation of ecotypes occurs instantaneously (<1000 generations) rather than gradually. Although ecotypes have formed, individuals in two ecotypes are still able to interbreed with each other even from different locations (Hollander et al. 2005). The survival rate of hybrids in the contact zone is higher than parental ecotypes or on the same level which means hybrid superiority exists (Rolan-Alvarez et al. 1997). It promotes the first step of speciation, ecotype formation because hybrids are easy to back-cross with parental species. However, it impedes the completion of speciation due to continuous gene flow.

Gene flow between crab and wave ecotypes never disappear although in the contact zone gene flow is impeded and less than gene flow within ecotypes (Panova et al. 2006). How *Littorina saxatilis* maintain divergence in the face of gene flow which is regarded as genetic recombination counteracting divergence? Strong enough selection pressure could overcome this kind of homogenizing effects and genomic architectures might resist gene flow by impeding gene recombination (Smadja & Butlin 2011). Hybrid zone analysis has inferred patterns of selection in space of *Littorina saxatilis* that crabs represent a strong selection pressure during ecotypes separation (Westram et al. 2018).The genetic basis of parallel adaptation is supposed to depend on underlying genomic architecture (Yeaman 2013) such as the number, size and additivity of loci and nonrandom arrangements order in the genome. Chromosome inversion, a kind of chromosome rearrangement, has been researched widely. It is considered to be able to suppress recombination and serve as reservoirs for adaptive standing variation which might be the reason for fast parallel adaptation (Morales et al. 2019). In addition, assortative mating of ecotypes is obvious in both field and laboratory experiments. Individuals are more likely to mate with the same ecotype individuals by means of following trails and longer mating time (Hollander et al. 2005). In brief, the partial speciation of this species is considered to be evolved by divergent selection of heterogeneous environments, habitat choice, assortative mating, and genomic background (chromosome rearrangement).

In our research site, Spanish, ecotypes are distributed over vertical shore gradients. Hybrids are a minority compared with parental ecotypes in the intermediate habitat (Johannesson et al. 1993) with isolation indexes of 0.5–0.9 in Spain tested from mate choice experiments (Johannesson et al. 1995). The Spanish system of *Littorina saxatilis* ecological divergence is found much older than Swedish and British systems based on mitochondrial DNA lineages (Panova et al. 2011). Furthermore, Morales (Morales et al. 2019) found that the comparisons between crab ecotype and wave ecotype showed more significant divergence in Spain than in Sweden. However, the

comparisons only focus on crab ecotype and wave ecotype snails themselves and ignore hybrids. Hybrid of two ecotypes could offer more information than distinct population comparisons. For instance, cline analysis and hybrid zone analysis in Sweden (Westram et al. 2018) were used to detect non-neutral SNP (single nucleotide polymorphism) and find a genotype-phenotype-environment association. Therefore, this project will take hybrid whole genome sequencing into account.

Then what kind of sequencing method and subsequent data analysis should be used in this research? Low coverage whole genome sequencing (lc WGS) is regarded as a cost-efficient approach to catch low frequency variation in many individual samples of population (Gilly et al. 2018). The development of calling algorithms makes it possible to call SNP and estimate allele frequency accurately from low depth sequencing data. Analysis of next generation sequencing data (ANGSD) is a software designed for next generation sequencing data with data analysis methods of taking genotype uncertainty into account instead of calling genotypes directly (Korneliussen et al. 2014). It is especially suitable and useful for low and medium depth data. Therefore, it is a good sequencing data analysis tool for lc WGS data of population genomics. ANGSD has been used to explore underlying genetic mechanisms of parallel phenotypic evolution, local adaptation, history of evolution, and so on (Wilder et al. 2020) (Therkildsen et al. 2019) based on population genetic sequencing data. Furthermore, linkage disequilibrium is a good indicator to explore population genetic history including selection, domestication, chromosome rearrangement, and so on. Also, LD decay analysis could reveal population recombination and selection history (Zhang et al. 2018). All of these measures help us to explore population structure, natural selection signature, and history of genetic changes.

In this research, I expect to find a genetic continuous cline like the Sweden system from principal components analysis (PCA), special population genetic differentiation, and nucleotide density pattern in chromosomes, detect new inversions candidates and infer parallel adaptive evolution history of *Littorina saxatilis* in Spain. Combining all above, I try to explain the underlying mechanism of parallel ecological divergence and the process of speciation facing gene flow.

2 Method

Samples collection, DNA sequencing, and processing were conducted by staff at the University of Sheffield. Also, details of these processes were provided by co-supervisor Francesca and Roger and described in the supplementary information.

2.1 PCA

After getting bam files, I used software ANGSD v.0.934 (Korneliussen et al. 2014) (http://www.popgen.dk/angsd/index.php/ANGSD) to estimate imputed genotype probabilities from mapped reads. ANGSD is a software designed for analyzing next generation sequencing data, especially low and medium depth data due to its feature of taking genotype uncertainty into account. In order to explore the population structure of Littorina saxatilis in Spain, I performed principal components analysis (PCA) and admixture analysis. At first, I used ANGSD to perform SNP calling, estimate genotype likelihood in beagle format and get a list of SNP positions as input files to estimate covariance matrix. Both ANGSD and PCAngsd could infer covariance matrix of Littorina saxatilis 73 samples. Here I chose PCAngsd v.1.02 (Meisner & Albrechtsen 2018) (http://www.popgen.dk/software/index.php/ PCAngsd) and then used R v.4.0.3 to perform PCA using eigen function and extract eigenvectors to plot PC1 vs PC2. I also used individuals' distance along crab-hybrid-wave axis data as color information to show population structure along crab-hybrid-wave axis.

2.2 Admixture analysis

Another method to interpret population structure is admixture analysis which showed genome-wide admixture proportions for every individual. ngsAdmix (http://www.popgen.dk/software/index.php/NgsAdmix) (Skotte et al. 2013)used beagle format genotype likelihood file to infer admixture proportions of ancestry clusters for every individual. Then I used R to draw admixture proportions bar plot along crab-hybrid-wave axis. After getting the admixture analysis result, I could classify 73 individuals into 3 groups: crab, hybrid and wave according to individual ancestry clusters contribution. Given ecotype information, I could draw a new PCA plot in 3 colors showing population structures of every ecotype population.

2.3 Folded one dimensional and two dimensional site frequency spectrum

In order to detect selection signatures on genome and understand patterns of divergence and differentiation across genome, I calculated several summary statistics from low-depth NGS data. The first step was estimating sample allele frequencies (SAF) posterior probabilities by ANGSD for each population separately. Then I used the program realSFS of ANGSD to calculate Site Frequency Spectrum (SFS) which records the proportions of sites at different allele frequencies. Here I calculated folded SFS without outgroup species defining ancestral state. After getting one dimensional folded SFS, I drew a bar plot for every population removing the first value of 1D folded SFS which represented the expected number of sites with derived allele frequency equal to 0 due to its relatively too big value to guarantee other values clear interpretation. The next

step was estimating joint SFS between 2 populations (2D folded SFS) which is useful to infer the divergence process of populations and estimate population genetic differentiation as prior information. As before calculating 1D folded SFS, I used program realSFS to infer 2D folded SFS between every pair of 3 ecotype populations. Similarly, I could use R to illustrate these flatten matrixes in heat map format.

2.4 Population genetic differentiation and nucleotide diversity

Here I chose fixation index (FST) (Hudson et al. 1992) and population branch statistic (PBS) as indicators of allele frequency differentiation. ANGSD helps us calculate FST and PBS from sample allele frequencies likelihoods (.saf files), avoiding genotype calling. The first step was computing per-site FST indexes by realSFS using SAF files and 2D folded SFS as input. The output file of this stage is (a) and (a+b) values calculated in the way of Reynolds(Reynolds et al. 1983) for three FST comparisons for every SNP site. I could use R to calculate FST value for every SNP and every pairwise of 3 populations and then drew FST in every chromosome and every pairwise populations in the order of position. The next step was performing a sliding-window analysis with setting window size as 10kb and step size as 1kb by realSFS. Similarly I drew window PBS values in every chromosome and population.

Furthermore, in order to know whether allele frequency differentiation increase is related to the change of nucleotide diversity (Korneliussen et al. 2013), I used realSFS saf2theta function and program thetaStat of ANGSD to calculate thetas for each site and in windows taking SAF and SFS as input files and prior information. Pairwise theta estimation (tP) divided by numbers of sites in the window (nSites) can be used to estimate window-based pairwise nucleotide diversity (pi).

2.5 Linkage disequilibrium heat map and decay analysis

Linkage disequilibrium is a useful indicator to illustrate the correlation of pairwise loci and identify inversions. Here are various different measures of LD such as pairwise r2, Pearson estimation of r2, D from EM algorithm and so on. Here I choose pairwise r2 (coefficient of correlation) because it is very useful when analyzing biallelic markers such as SNPs and independent of sample size (Devlin and Risch 1995). I subset beagle files and maf files of every chromosome at first, and then sample these SNP sites to guarantee the total SNP number of every chromosome is around 10000 as instruction of ngsLD software. ngsLD v.1.1.1 (Fox et al. 2019)(https://github.com/fgvieira/ngsLD) is a software designed for NGS data taking the uncertainty of genotype's assignation into account. After using this software calculating LD among SNPs in every chromosome, I used R to draw LD heat map in whole chromosome and partial chromosome inversion candidates. Then I used a modified R script offered by ngsLD to analyze LD decay and draw LD decay line for every chromosome and population.

3 Results

3.1 PCA

The covariance matrix of principal components analysis (PCA) could be estimated by single-read sampling in ANGSD. With proper parameters settings, there are 12,817,872 variant sites (SNP) called from 73 individuals' bam files. Here I used PCAngsd to take genotype likelihoods in beagle format as input and then infer covariance matrix which is used in Figure 1.

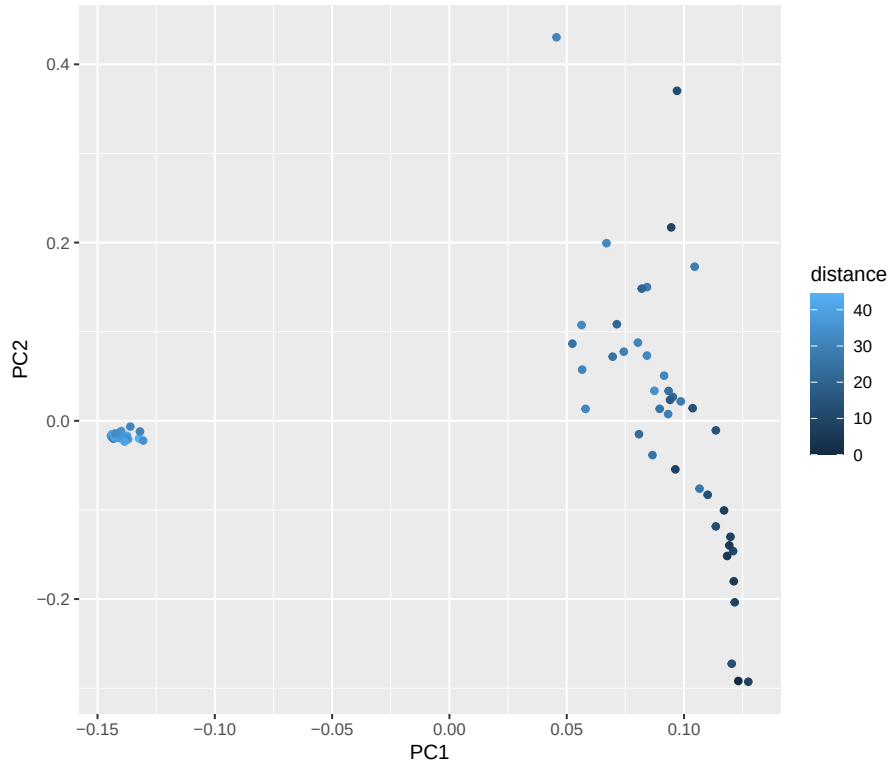

Figure 1: **PCA of 73 Littorina saxatilis individuals**
Dark blue presents crab ecotype and light blue presents wave ecotype. The more dark means the sampling individual is closer to crab ecotype along the crab-hybrid-wave axis. V2 is the first principal component and V3 is the second principal component.

The PCA result was quite different from the Sweden system. There are many individuals points squeezing together in a little circle region and the color of these individuals is light blue which means these individuals are likely to be wave ecotype snails. On the other side of Figure 1, points are darker and comparatively scattering. Furthermore, deep dark blue points are usually lower than medium blue points. So far, I could only know that there are 2 genetic distinct clusters along crab-hybrid-wave axis.

3.2 Admixture analysis

In order to infer every individual's genome-wide admixture proportions, I used ngsAdmix to estimate admixture proportions from genotype likelihoods and ancestry clusters number setting.

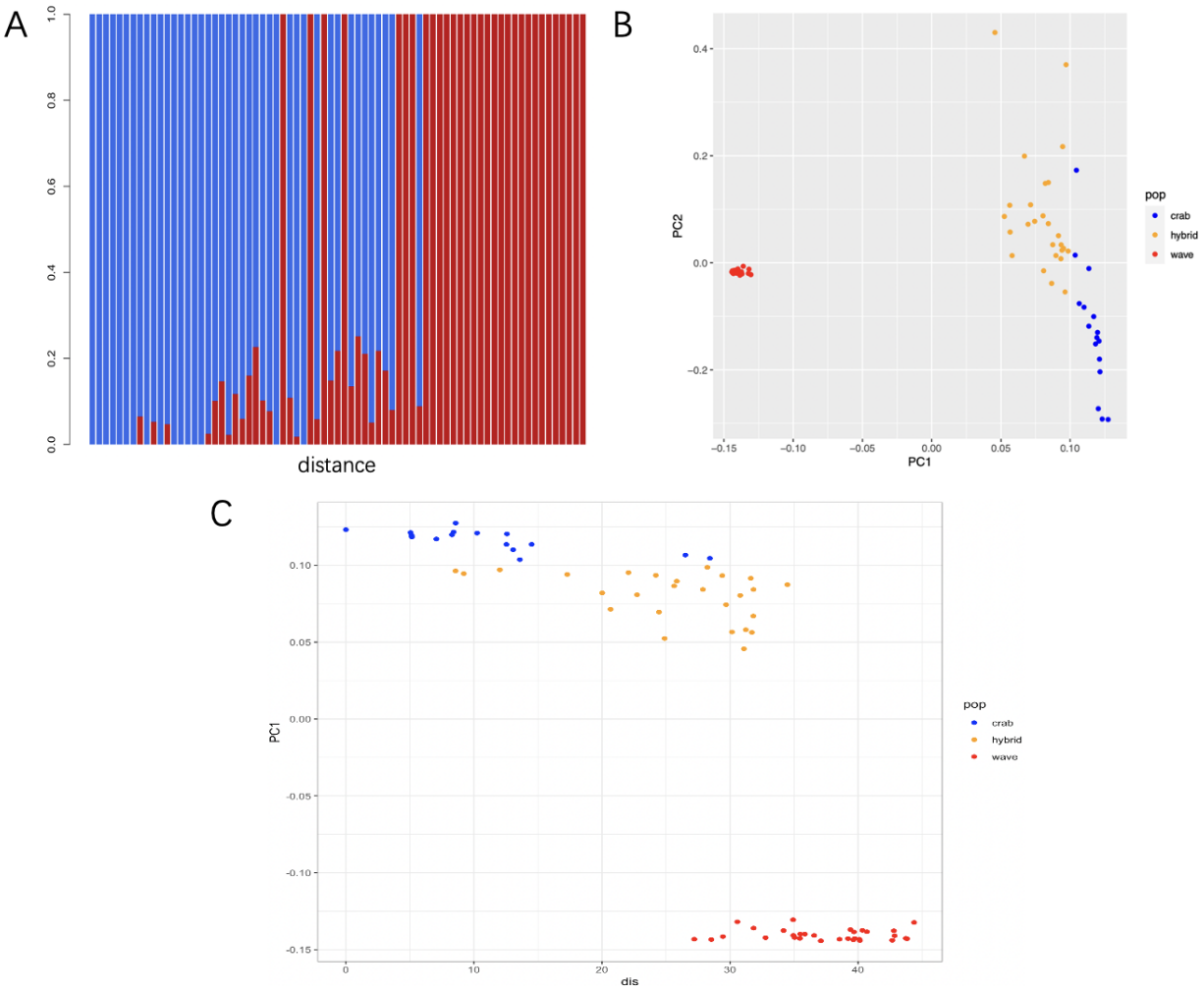

Figure 2: **Admixture analysis of 2 ancestry clusters and new PCA based on admixture analysis**

A. The figure showed individuals admixture analysis results along crab-hybrid-wave axis from crab to wave ecotype. Blue represents crab ecotype and red represents wave ecotype.

B. Based on admixture analysis, individuals could be divided into 3 groups: wave, hybrid and crab. Given the group information, PC1 vs PC2 of principal components analysis is shown in 3 colors.

C. Based on admixture analysis, 3 groups named wave, hybrid and crab could be plotted in a PCA between distance of crab-hybrid-wave axis and PC1.

As Figure 2A showed, right individuals are mostly wave ecotypes while left individuals are mostly crab ecotypes which corresponds with distance data along crab-hybrid-wave axis. However, in the middle part of Figure 2A, there are crab ecotype, wave ecotype and hybrid individuals, which means in the contact zone, 2 ecotypes and their hybrid offspring are mixed. Furthermore, hybrid individuals can also enter pure wave and pure crab ecotype zone. From the result of admixture analysis, I could divide individuals into 3 groups: crab ecotype, wave ecotype and hybrid with 16, 31 and 26 individuals separately. Hybrids include all individuals with admixture proportions that are not 0 or 100%. When group information is considered, PCA plot could be redrawn as Figure 2B. All of wave ecotype individual points squeezed together just as Figure 1. Crab ecotype individuals and hybrid individuals are all on the other side, far away from wave ecotype, which means crab group and hybrid group are much more close compared with wave group. Meanwhile, the hybrid group is higher in the second principal component than crab group, although it is difficult to distinguish hybrid group and crab group only using the first two principal components.

In Figure 2C, all wave individuals are around -0.14 in PC1, all crab individuals are higher than 0.10 while hybrid individuals are lower than 0.10 in PC1. However, only using the line of PC1=0.10 is not sufficient to split the higher cluster into crab and hybrid two isolated populations. Therefore, I would like to regard that there are 2 distinct groups, one is a cline within crab and hybrid group, the other is wave group.

3.3 1d SFS

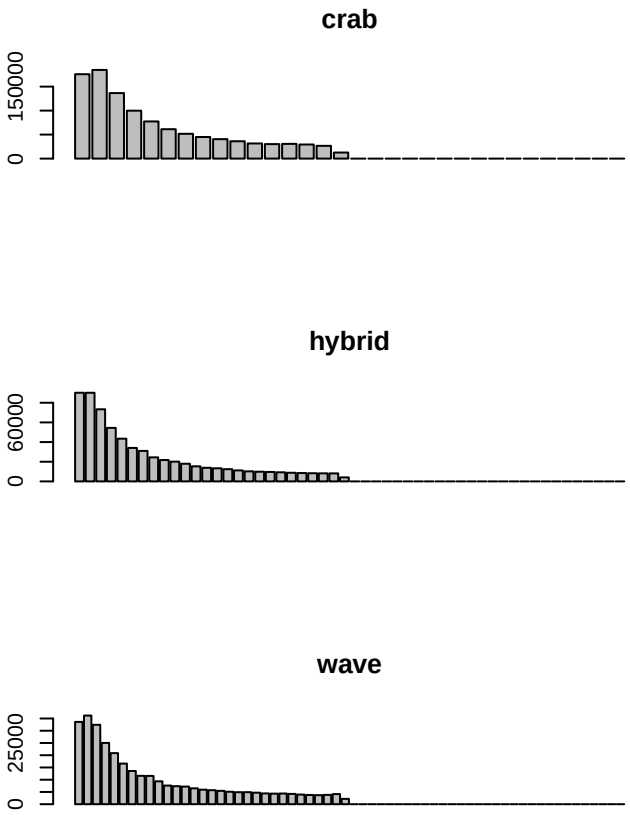

Figure 3: **One-dimensional folded Site Frequency Spectrum (SFS)**
One-dimensional folded SFS of each group is shown in 3 bar plots from allele frequency 1 to the number of individuals in the group

Site frequency spectrum records the number of sites at different allele frequencies. One-dimensional folded SFS only contains allele frequencies in one group. In Figure 3, sites will be recorded in SFS only if this site is caught in every individual of the group, also I have removed all of the sites without alleles due to its comparatively huge value. As the bar plots showed, there are fewer sites of allele frequency 1 than allele frequency 2 in crab and wave group. It is unexpected but it might be caused by distortion effect on SFS of polymorphic inversions and SNP filter conditions. The sites number of crab group is bigger than hybrid and wave group. It could be explained by smaller population size of crab group which means it is easier for crab group to detect the same site in every individual. Furthermore, after calculating the proportion of SNP sites in every group, crab group SNP proportion is 0.0193, wave group SNP proportion is 0.0185, hybrid group SNP proportion is 0.02266. The hybrid group has the highest SNP proportion maybe because the hybrid group contains SNP both in crab group and wave group.

3.4 2d SFS

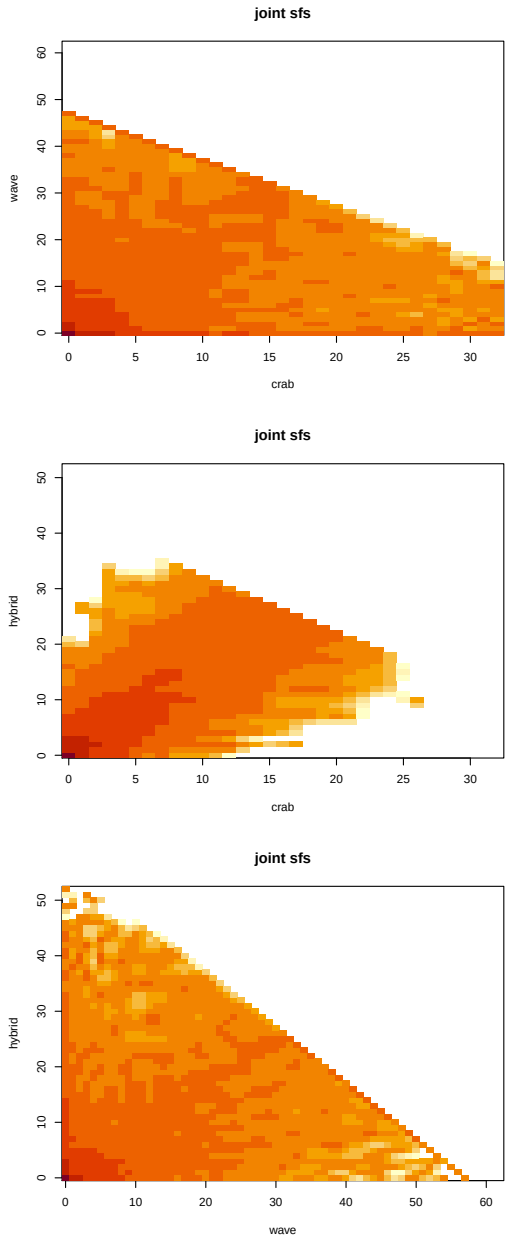

Figure 4: **Two-dimensional folded Site Frequency Spectrum (SFS)**
Two-dimensional folded SFS of each pair of groups are shown in 3 heat plots from allele frequency 0. Darker color means more sites and lighter color means less sites.

Joint SFS between 2 populations could be used to infer the divergence process. From Figure 4, it is clear that crab and hybrid groups are closer because there are darker color squares on the diagonal compared with the other 2 heat plots. Usually colored squares of folded 2D-SFS are all under one of the diagonals. However in Figure 4, 2D folded SFS didn't meet the expectation. This situation might be resulted from different population sizes. 2D-SFS could also be used as prior information for estimating FST and PBS (population branch statistic) which will be discussed later.

3.5 Summary statistics

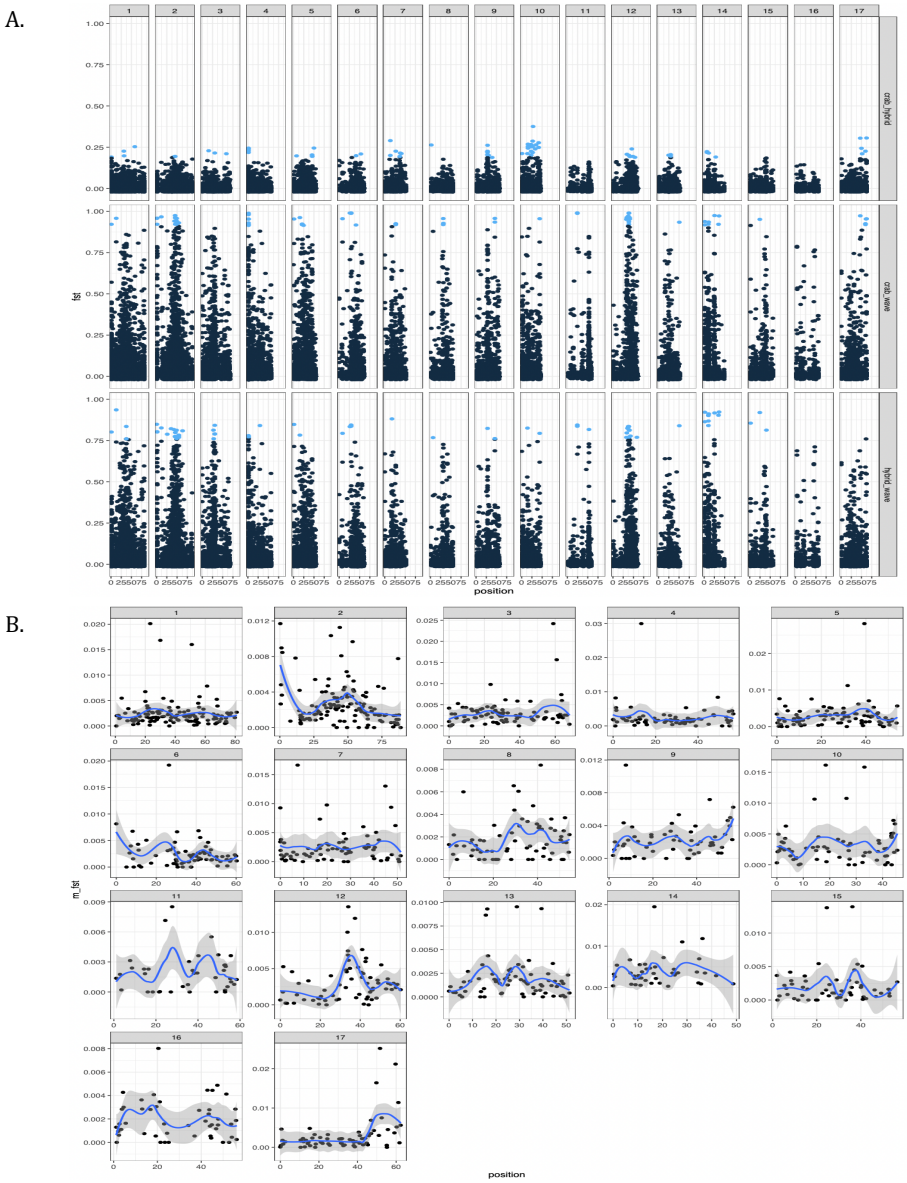

Figure 5: **FST value among 17 chromosome and three pairwise populations**
A. FST value scatter plot of every chromosome and pairwise populations with blue points representing outliers.
B. Mean FST value between crab and wave ecotype in every position on chromosome along with smooth lines showing trend.

There are some big blocks showing high FST value, high PBS value and low pi value, which are all signatures of positive selection, especially in LG 1, 2, 4, 6, 9, 12, 14, 17. From Figure 5A, I can find that population genetic differentiation of crab and wave is the largest. Also, it is clear that hybrid and crab are closer due to their low FST value in all chromosomes. Figure 5B exhibits mean FST value of crab and wave population showing general high FST value in the specific position such as crest part of LG2, 6, 12, 17. Population branch statistics of hybrid is the lowest which is not difficult to understand because hybrids contain genetic components from

crab and wave ecotype. PBS of wave is the highest which is another proof of closer relationship between crab and hybrid. Pairwise nucleotide in Figure 2 in supplementary information can not offer too much effective information. The pi value looks nearly all the same in every chromosome and every ecotype.

Faria et al. (Faria et al. 2019) found some chromosome inversion candidates as Figure 6A shown. In these regions, FST and PBS are usually relatively higher than other regions, while pi are usually relatively lower than other regions. In order to interpret differences between these chromosomal rearrangement candidates and other regions, I can also calculate mean value of FST, PBS, pi and linkage disequilibrium (LD, which will be discussed later) in these regions and whole chromosome. Table S1 in the supplementary information gives the exact number of these statistics values which is more helpful to some extent. Also, a scatter plot Figure 6B has been drawn to interpret differences between candidates and whole chromosome more intuitively. From Figure 6B, generally FST and LD values of chromosome inversion candidate regions are higher than whole chromosome. However, PBS and pi value of chromosome inversion candidates region is sometimes lower and sometimes higher than whole chromosome. Statistics of LG1.2, 2.1, 4.1, 6.1, 6.2, 9.1, 14.3 meet the expected result of higher PBS and lower pi in inversion candidates regions. LG1.1, 7.1, 7.2, 17.1 show higher PBS and higher pi while LG 10.1, 11.1, 12.1, 12.2, 14.1, 14.2 show lower PBS and higher pi.

However, there are also some chromosome regions showing high FST and PBS and low pi, but are not chromosomal rearrangement candidates, such as the middle part of LG 2, 8, 11 and 15. They will be discussed later.

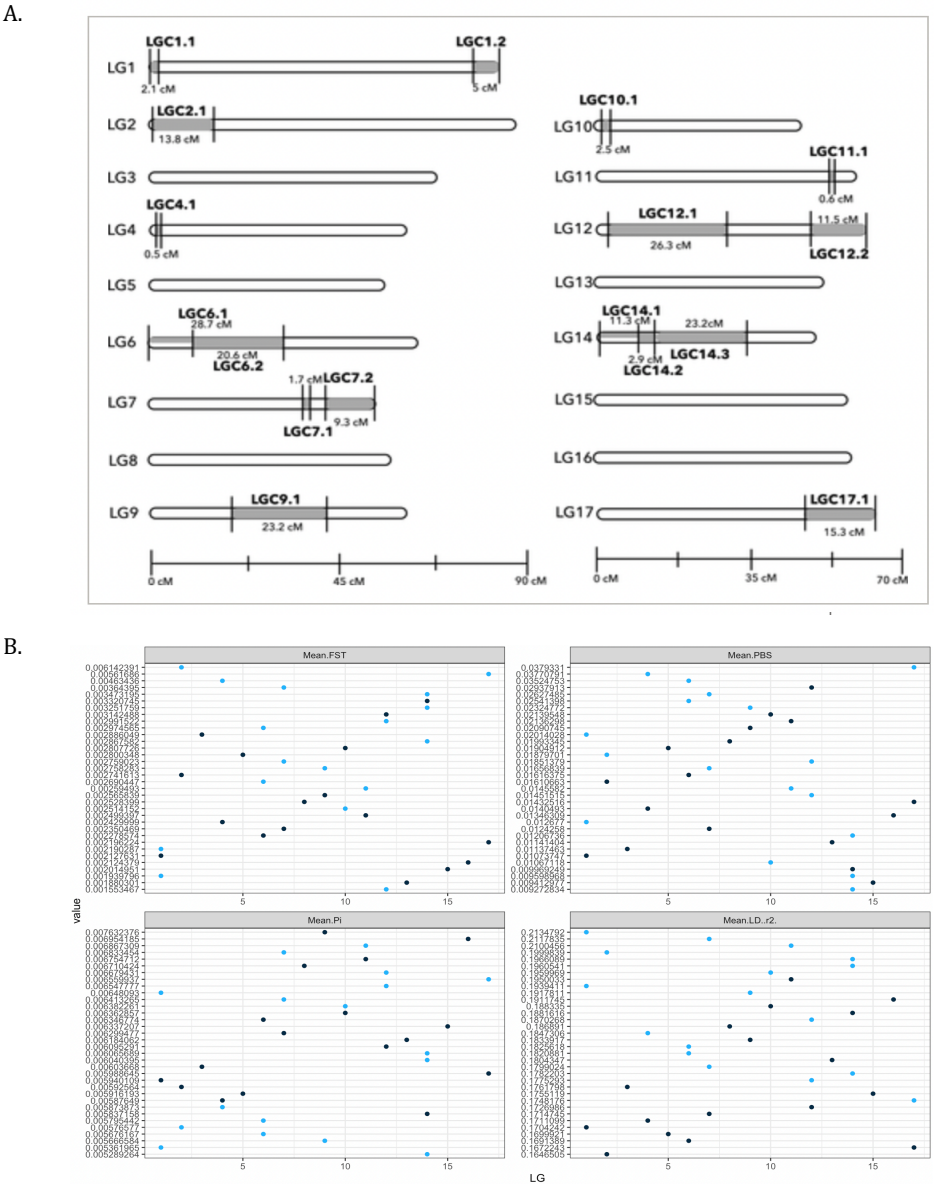

Figure 6: **Mean statistics of inversion candidates and whole chromosome**

A. Chromosomal rearrangement candidates inferred by Faria et al.(Faria et al. 2019)

B. Blue represents inversion candidates mean statistics while black represents whole chromosome mean statistics. The statistics here include mean FST of crab and wave ecotype, mean PBS of crab ecotype, mean pi of crab ecotype and mean LD of crab ecotype.

3.6 Linkage disequilibrium

Linkage disequilibrium estimates the nonrandom association of alleles at different loci. Natural selection, mutation, non-random mating, gene flow and population size can all affect LD. Here I choose pairwise r2 (coefficient of correlation) because it is very useful when analyzing biallelic markers such as SNPs and independent of sample size (Devlin and Risch 1995). All of the LD heat map and LD decay analysis regression plots are included in Figure 3 and 4 of supplementary information. Here Figure 7 only gives an example on chromosome 14. In 17 chromosomes, the crab population always has the highest LD value. Hybrid usually has medium LD value and wave has the lowest LD value.

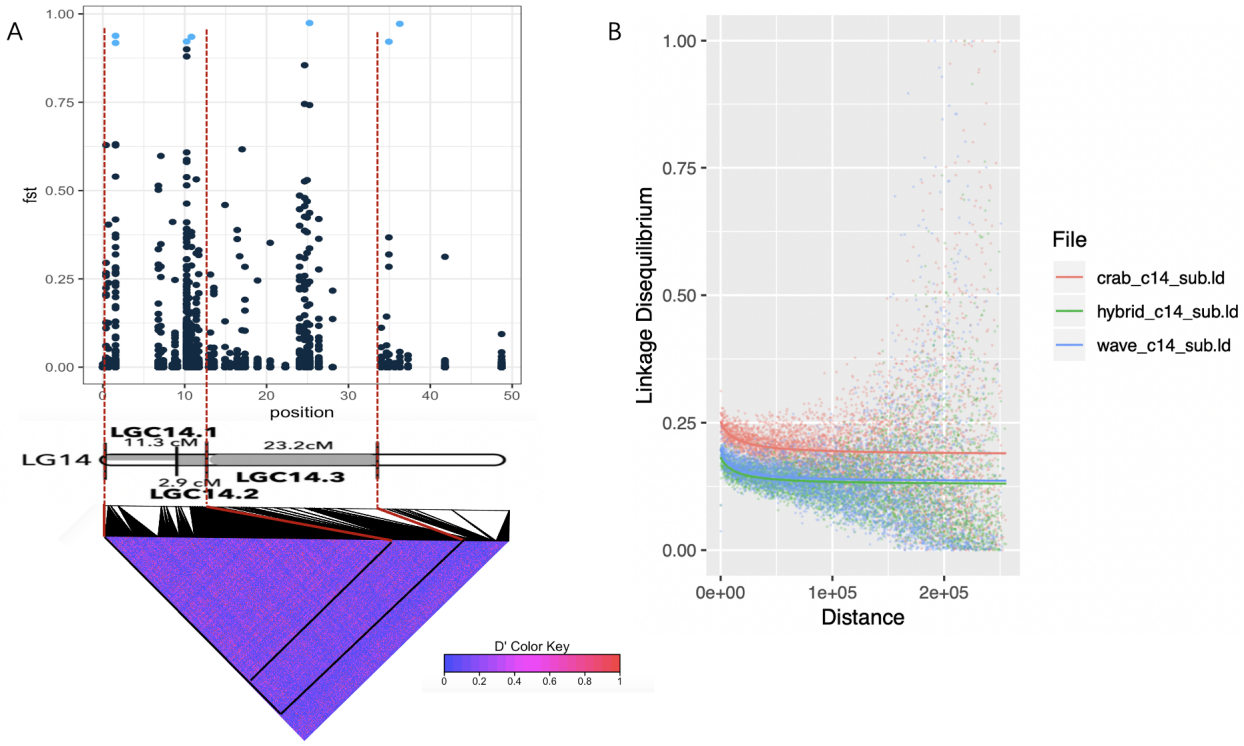

Figure 7: **LD heat map and LD decay of chromosome 14**
A. FST value with outliers mapping to chromosome inversion candidates and LD heatmap of crab population on chromosome 14.
B. LD decay analysis of chromosome 14

LD heat map among the whole chromosome is hard to get useful information because usually all regions look the same. However in LG 14, the heat map between the first two red dotted lines has a clear higher LD value block which maps to LG14.1 and LG14.2. The region between the second and the third red dotted lines is also more pink than white region on chromosome 14. Partial LD heat map in chromosome inversion candidates is more clearly showing higher overall LD value compared with other regions. Exact mean LD values are shown in Table S1 that mean LD values of chromosomal rearrangement candidates are higher than other regions.

For LD decay analysis, every single point is mean LD value of a 50bp size window. The fitting line decreases quickly approaching a horizontal line. As the distance increases, points become

more dispersed and some chromosomes disperse more early such as LG1 and LG4. Crab usually has the slowest LD decay rate in 17 chromosomes while hybrid ecotype population LD decay rate is usually the fastest, which might be resulted from higher population genetic diversity gained from crab and wave ecotype population.

4 Discussion

4.1 PCA and admixture analysis

There are 2 distinct clusters in PCA in the Spain system instead of a cline which might be resulted from more time to accumulate divergence. The first group is more homogeneous and regarded as wave ecotype. The second group is more heterogeneous and regarded as a cline within crab and hybrid group. Furthermore, from the admixture analysis, I could find obvious directional introgression from wave into crab at a low rate.

In Sweden admixture analysis conducted by Sheffield University staff, there were some individuals showing higher than 50 % of wave contribution, but in the older system of Spain, the situation is different. A possible speculation is that, in the process of speciation, hybrids are more likely to mate with crab (or to say bigger individuals), then hybrid population in PCA become closer and closer to crab population, ancestry proportion of wave decrease and proportion of crab increase. In addition to assortative mating, the possible reason for one direction of introgression could also be asymmetrical selection, or confusion of adaptation to the lower shore within crab group or a mix of these.

4.2 Summary statistics of differentiation

Here I expect to find special population genetic differentiation and nucleotide density pattern in chromosomes, detect new inversions candidates and infer parallel adaptive evolution history.

4.2.1 SFS

The 1d SFS seems to be far from the neutral expectation. It could not be caused by introgression since it is similar in all groups. The SFS might be distorted by SNPs in polymorphic inversions. I only used sites detected in all individuals to guarantee the accuracy of gene frequency, but this might also distort the SFS. For instance, if repeat regions are more likely to fall into this class, the SFS will be inaccurate. From 2d SFS, the only significant information is that crab and hybrid populations are closer which is consistent with PCA result.

4.2.2 Population genetic differentiation

Most of loci FST value doesn't reach 1 which means these loci are not fixed and keep polymorphic. Rare fixed loci suggest balancing selection which tends to prevent fixation. The same pattern for SNPs was found by Westram (Westram et al. 2018) and some possible explanations were given: indirect divergent selection on SNPs linked to selected variants, selection on polygenic traits or a combination of divergent selection and balancing selection that make loci maintain polymorphism in one or two habitat ends.

High FST means high divergence. Regions with lots of high FST values may well contain genes under divergent selection between crab and wave ecotypes although along with some noises. It is

hard to find these regions just from FST scatter plot due to some regions' more SNPs. Usually, regions with many SNPs tend to own high FST SNPs. Therefore, I colored the top 0.01% of SNPs as Figure5A and calculated mean FST per map position showed as Figure5B. There are some of the same regions detected before by Morales (Morales et al. 2019) including inversion candidates. There are also some regions on LG2, 8, 11 demonstrating high population genetic differentiation but not regarded as inversion candidates. However, further exploration should be executed to verify whether they are new inversions or just genetic speciation islands that harbor loci underlying reproductive isolation, or regions of low recombination.

Generally, I expect to find high FST, high PBS, low pi and high LD regions as candidate blocks for reproductive isolation. Chromosome inversion candidates are supposed to carry many loci under divergent selection, so I expect to detect the same summary statistics pattern in these candidates. However, I found that many candidates show higher pi, even lower FST and PBS. Only LG1.2, 2.1, 4.1, 6.1, 6.2, 9.1 meet all conditions. Higher pi is possible when inversions are polymorphic. The elevated diversity could be between arrangements or within an arrangement at the nucleotide level due to balancing selection. For example, LG10.1, 12.1, 12.2, 14.1, 14.2 show lower mean FST, lower mean PBS but higher mean pi and higher mean LD. Balancing selection may also result in lower FST and lower PBS, but these regions still maintain high LD. Some inversion regions have similar levels of diversity to whole chromosome suggesting non-polymorphic and not accumulated differentiation. Inversion regions have less diversity suggesting these could have swept to high frequency recently.

It is worth noting that chromosome inversions can't fully explain parallelism. Morales also found shared outlier loci distributing across the genome (Morales et al. 2019). This could be caused by polygenic selection of multiple loci with small effect underlying parallel adaptive divergence.

4.2.3 Limitation of FST

In genomic regions surrounding barrier loci, the barrier effect initially allows a build-up of genetic differentiation in the form of allele frequency variation between populations, typically using a relative measure such as FST (Ravinet et al. 2017). However, Using FST as an indicator of divergent selection has some problems.

The pattern of FST is affected by multiple factors that change throughout the genome, including mutation, genetic drift, selection, demographic history, recombination, gene flow, gene density, and genome structure, some of which are expected to change at different stages of speciation. Polygenic makes it possible that FST of many underlying loci remain low when snails achieving trait divergence(Westram et al. 2014). Background and positive selection may produce similar patterns in genome scan using FST, reducing intraspecific diversity and increasing interspecific gene composition(Cruickshank & Hahn 2014). Furthermore, some non-inversion candidates regions with high FST may not be other currently undetected chromosomal inversions under positive divergence selection, but related to other mechanisms suppressing recombination such

as centromere region combined with background selection or divergence hitchhiking (Ravinet et al. 2017).

4.3 Linkage disequilibrium

From LD analysis, I could find that in 17 chromosomes, the crab population always has the highest LD value. LD is likely to be high in crab for many reasons. It might be because crab has a small effective population size, or crab suffered bottleneck accident in history, or crab is under a relatively strong selection, or crab individuals have a stronger tend of assortative mating. Genetic structure may also increase LD. Polymorphic inversions are able to suppress recombination. The clines generate LD whether they are due to introgression or due to a selective gradient. Generally, hybrid has medium LD value and wave has the lowest LD value, but the difference is little. The selection of wave environment is not as strong as crab, so it is reasonable that LD value of wave ecotype population is the lowest, and hybrid is medium. The little LD difference between hybrid and wave groups indicates that the directional introgression destroys high LD genetic structure in crab group.

LD heat map should reveal inversions and a general pattern of higher LD among neighboring loci, especially among SNPs in the same contig. However, in LD heat map of 17 chromosomes in *Littorina saxatilis*, there is usually no significant higher LD value block. Chromosome inversion candidates LG4.1, 7.2, 9.1, 11.1, 14.1, 14.2 are relatively clear to find in whole chromosome LD heat map compared with other inversion candidates. Therefore, it is more difficult to infer new inversion candidates just from LD heatmap.

For LD decay analysis, crab usually has the slowest LD decay rate in 17 chromosomes which means the selection effect on crab ecotype population is stronger than other populations. Selection will decrease population genetic diversity and strengthen association between pairwise loci (LD). Therefore, usually population under stronger selection has a slower LD decay rate. Littorina is a high fertility and short generation interval species, therefore LD decay rate in this species is generally fast. Hybrid ecotype population LD decay rate is usually the fastest which might be resulted from higher population genetic diversity gained from crab and wave ecotype populations.

Points disperse along the distance may be resulted from polymorphic inversions on chromosomes which makes high LD between SNPs far apart. In some chromosomes, points disperse more early than others which might be because short inversions on chromosome which result in LD between loci far away. The most obvious effect of inversions on LD decay is typically going to be on larger scales especially between breakpoints far apart (such as 10Mb), but due to the limitation of contig size, I could not find this situation.

Ravinet (Ravinet et al. 2016)found that LD clusters showed strong signals of ecotype specificity, loci were largely homozygous in crab while heterozygous in wave, suggesting these loci have

undergone selective sweeps (the reduction or elimination of variation at sites that are physically linked to a site under directional selection) in crab population. This is another proof that crab group suffered stronger selection fueling divergence.

4.4 Other possible mechanisms of parallel adaptation

Morales (Morales et al. 2019) supposed that standing variation within chromosomal inversions can be maintained as balanced polymorphism and fuel rapid parallel phenotypic divergence to heterogeneous environments through gene flow without high sharing of genomic outliers. Balancing and divergent selection between habitats could maintain inversions for long periods of time, resulting in the high diversity and divergence for some inversions noted above (Faria et al. 2019). Furthermore, balancing selection is often documented for inversion polymorphisms (Wellenreuther & Bernatchez 2018). Inversions can extend the impact of barrier loci (reproductive isolation loci) to linked loci and promote additional barrier loci accumulating in inverted regions with gene flow between populations which facilitates the efficient spread of adaptive standing variance in inversions (Morjan & Rieseberg 2004). After the establishment of local LD, gene drift can facilitate disruptive selection (Dieckmann & Doebeli 1999) which suggests random process may play a role in speciation.

There are some studies revealing that a large scale of SNPs for some key adaptive phenotypes or environmental axes could be explained by SNPs in chromosome inversions (Morales et al. 2019). Westram also found that 75% of non-neutral SNPs cluster together in three LGs (6, 14, 17) which suggests that large regions of low recombination are consistent with the presence of chromosomal rearrangements.

4.5 Future direction

From a speciation perspective, the main goal is to infer the number, distribution and intensity of gene flow barriers, and their impact on other genomic regions, then find genetic speciation island which harbor loci underlying reproductive isolation. However, due to many other processes' existence, it is challenging to detect this signal from genome scan. Chromosome inversions are regarded as candidates for gene flow barriers. Then the research on ages and spatial distributions of inversions will be important to understand local adaptation and evolution history of reproductive isolation in *Littorina saxatilis*. It is premature to use diversity and divergence data of *Littorina saxatilis* to calculate exact inversion ages, but Morals (Morales et al. 2019) inferred from Sweden snails data that the origins of inversions are earlier than postglacial colonization of the Swedish coast.

After finding inversions regions under divergence selection, the next step is to identify why they are so differentiated. These regions are likely to contain reproductive isolation loci which need more exploration, such as finding break point of inversions, functional annotation of outlier loci, selection scans, phenotype-habitat-genotype associations, experimental analyses with sequencing (Koch et al. 2021). It is worth mentioning that size is the feature under strong divergence

selection and also could be the main reason for assortative mating between ecotypes (Smadja & Butlin 2011), therefore size is likely to facilitate the formation of barriers between ecotypes. Then finding the association between size and reproductive isolation which is deduced by Johannesson (Johannesson 2016) is another promising research direction.

5 Conclusion

In general, I found two distinct groups in the Spanish system of *Littorina saxatilis* instead of a cline displayed in the more recent system Sweden. Compared with the wave ecotype, hybrids are much closer to the crab ecotype. Besides, the directional introgression from wave to crab destroys high LD in the crab ecotype. Differentiation analysis hints at balancing selection on this species and LD analysis hints at stronger selection and more polymorphic inversions of crab ecotype. Also, I found some new chromosome inversions candidates for future verification. This research tried to infer the process and mechanisms of *Littorina saxatilis* ecological divergence and speciation after the phase of the Sweden system. However, to explore the underlying mechanisms of parallel ecological divergence with gene flow and the evolvement of reproductive isolation with interbreeding, more efforts are still needed in future.

Code availability

All of the scripts, software parameter settings, and dependancies packages are included in the following website: https://github.com/rz520/LittorinaPipeline

Acknowledgement

Here I would like to appreciate supervisor Matteo Fumagalli for his continuous help on software usage, parameter setting, and results interpretation. Co-supervisor Francesca Raffini and Roger K. Butlin gave me lots of guidance on *Littorina saxatilis* background knowledge and recent research progress. All of my supervisors helped me to polish my thesis by giving constructive suggestion. I am also grateful to other master students in Matteo's Lab and my friends for their help, encouragement and company.

References

Beaumont, M. A. (2010), 'Approximate bayesian computation in evolution and ecology', *Annual Review of Ecology, Evolution, and Systematics* **41**(1), 379–406.
URL: *https://doi.org/10.1146/annurev-ecolsys-102209-144621*

Butlin, R. K., Galindo, J. & Grahame, J. W. (2008), 'Sympatric, parapatric or allopatric: the most important way to classify speciation?', *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1506), 2997–3007.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2008.0076*

Cruickshank, T. E. & Hahn, M. W. (2014), 'Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow', *Molecular Ecology* **23**(13), 3133–3157.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12796*

Dieckmann, U. & Doebeli, M. (1999), On the origin of species by sympatric speciation, Iiasa interim report, IIASA, Laxenburg, Austria.
URL: *http://pure.iiasa.ac.at/id/eprint/5926/*

Faria, R., Chaube, P., Morales, H. E., Larsson, T., Lemmon, A. R., Lemmon, E. M., Rafajlović, M., Panova, M., Ravinet, M., Johannesson, K., Westram, A. M. & Butlin, R. K. (2019), 'Multiple chromosomal rearrangements in a hybrid zone between littorina saxatilis ecotypes', *Molecular Ecology* **28**(6), 1375–1393.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14972*

Fitzpatrick, B. M., Fordyce, J. A. & Gavrilets, S. (2008), 'What, if anything, is sympatric speciation?', *Journal of Evolutionary Biology* **21**(6), 1452–1459.
URL: *https://doi.org/10.1111/j.1420-9101.2008.01611.x*

Fox, E. A., Wright, A. E., Fumagalli, M. & Vieira, F. G. (2019), 'ngsLD: evaluating linkage disequilibrium using genotype likelihoods', *Bioinformatics* **35**(19), 3855–3856.
URL: *https://doi.org/10.1093/bioinformatics/btz200*

Galindo, J., Martínez-Fernández, M., Rodríguez-Ramilo, S. T. & Rolán-Alvarez, E. (2013), 'The role of local ecology during hybridization at the initial stages of ecological speciation in a marine snail', *Journal of Evolutionary Biology* **26**(7), 1472–1487.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/jeb.12152*

Galindo, J., Morán, P. & Rolán-Alvarez, E. (2009), 'Comparing geographical genetic differentiation between candidate and noncandidate loci for adaptation strengthens support for parallel ecological divergence in the marine snail littorina saxatilis', *Molecular Ecology* **18**(5), 919–930.
URL: *https://doi.org/10.1111/j.1365-294X.2008.04076.x*

Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R., Melloni, G. E. M., Hatzikotoulas, K., Farmaki, A.-E., Ritchie, G., Schwartzentruber, J., Danecek, P., Kilian, B., Pollard,

M. O., Ge, X., Tsafantakis, E., Dedoussis, G. & Zeggini, E. (2018), 'Very low-depth whole-genome sequencing in complex trait association studies', *Bioinformatics* **35**(15), 2555–2561.
URL: *https://doi.org/10.1093/bioinformatics/bty1032*

Grahame, J. W., Wilding, C. S. & Butlin, R. K. (2006), 'Adaptation to a steep environmental gradient and an associated barrier to gene exchange in littorina saxatilis', *Evolution* **60**(2), 268–278.
URL: *https://doi.org/10.1111/j.0014-3820.2006.tb01105.x*

Hollander, J., Lindegarth, M. & Johannesson, K. (2005), 'Local adaptation but not geographical separation promotes assortative mating in a snail', *Animal Behaviour* **70**(5), 1209–1219.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347205002733*

Hudson, R. R., Slatkin, M. & Maddison, W. P. (1992), 'Estimation of levels of gene flow from dna sequence data.', *Genetics* **132**(2), 583–589.
URL: *https://www.genetics.org/content/132/2/583*

Johannesson, K. (2016), 'What can be learnt from a snail?', *Evolutionary Applications* **9**(1), 153–165.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.12277*

Johannesson, K., Johannesson, B. & Rolán-Alvarez, E. (1993), 'Morphological differentiation and genetic cohesiveness over a microenvironmental gradient in the marine snail littorina saxatilis', *Evolution* **47**(6), 1770–1787.
URL: *https://doi.org/10.1111/j.1558-5646.1993.tb01268.x*

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E. & Butlin, R. K. (2010), 'Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation', *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1547), 1735–1747.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2009.0256*

Johannesson, K., Rolán-Alvarez, E. & Ekendahl, A. (1995), 'Incipient reproductive isolation between two sympatric morphs of the intertidal snail littorina saxatilis', *Evolution* **49**(6), 1180–1190.
URL: *https://doi.org/10.1111/j.1558-5646.1995.tb04445.x*

Kirkpatrick, M. & Ravigné, V. (2002), 'Speciation by natural and sexual selection: Models and experiments', *The American Naturalist* **159**(S3), S22–S35.
URL: *http://www.jstor.org/stable/10.1086/338370*

Koch, A., Brierley, C. & Lewis, S. L. (2021), 'Effects of earth system feedbacks on the potential mitigation of large-scale tropical forest restoration', *Biogeosciences* **18**(8), 2627–2647.
URL: *https://bg.copernicus.org/articles/18/2627/2021/*

Korneliussen, T., Moltke, I., Albrechtsen, A. & Nielsen, R. (2013), 'Calculation of tajima's d and other neutrality test statistics from low depth next-generation sequencing data', *BMC*

Bioinformatics **14**(1), 289.
URL: *http://www.biomedcentral.com/1471-2105/14/289*

Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. (2014), 'ANGSD: Analysis of next generation sequencing data', *BMC Bioinformatics* **15**(1), 356.
URL: *http://www.biomedcentral.com/1471-2105/15/356/abstract*

Meisner, J. & Albrechtsen, A. (2018), 'Inferring population structure and admixture proportions in low-depth ngs data', *Genetics* **210**(2), 719–731.
URL: *https://www.genetics.org/content/210/2/719*

Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M. & Butlin, R. K. (2019), 'Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast', *Science Advances* **5**(12).
URL: *https://advances.sciencemag.org/content/5/12/eaav9963*

Morjan, C. L. & Rieseberg, L. H. (2004), 'How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles', *Molecular Ecology* **13**(6), 1341–1356.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2004.02164.x*

Panova, M., Aronsson, H., Cameron, R., Dahl, P., Godhe, A., Lind, U., Ortega-Martinez, O., Pereyra, R., Tesson, S., Wrange, A.-L., Blomberg, A. & Johannesson, K. (2016), *DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms*, Vol. 1452, pp. 13–44.
URL: *https://pubmed.ncbi.nlm.nih.gov/27460368/*

Panova, M., Blakeslee, A. M. H., Miller, A. W., Mäkinen, T., Ruiz, G. M., Johannesson, K. & André, C. (2011), 'Glacial history of the north atlantic marine snail, littorina saxatilis, inferred from distribution of mitochondrial dna lineages', *PLOS ONE* **6**(3), 1–14.
URL: *https://doi.org/10.1371/journal.pone.0017511*

Panova, M., Hollander, J. & Johannesson, K. (2006), 'Site-specific genetic divergence in parallel hybrid zones suggests nonallopatric evolution of reproductive barriers', *Molecular Ecology* **15**(13), 4021–4031.
URL: *https://doi.org/10.1111/j.1365-294X.2006.03067.x*

Rafajlovic, M., Eriksson, A., Rimark, A., Hintz-Saltin, S., Charrier, G., Panova, M., Andre, C., Johannesson, K. & Mehlig, B. (2013), 'The Effect of Multiple Paternity on Genetic Diversity of Small Populations during and after Colonisation', *PLOS ONE* **8**(10).
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075587*

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B. & Westram, A. M. (2017), 'Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow', *Journal of Evolutionary Biology* **30**(8), 1450–1477.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/jeb.13047*

Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C. & Panova, M. (2016), 'Shared and nonshared genomic divergence in parallel ecotypes of littorina saxatilis at a local scale', *Molecular Ecology* **25**(1), 287–305.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13332*

Reid, D. G. (1996), *Systematics and evolution of Littorina*, Ray Society.
URL: *https://www.nhbs.com/systematics-and-evolution-of-littorina-book*

Reynolds, J., Weir, B. S. & Cockerham, C. C. (1983), 'Estimation of the coancestry coefficient: Basis for a short-term genetic distance', *Genetics* **105**(3), 767–779.
URL: *https://www.genetics.org/content/105/3/767*

Rolan-Alvarez, E., Johannesson, K. & Erlandsson, J. (1997), 'The maintenance of a cline in the marine snail Littorina saxatilis: The role of home site advantage and hybrid fitness', *EVOLUTION* **51**(6), 1838–1847.
URL: *https://doi.org/10.1111/j.1558-5646.1997.tb05107.x*

Schluter, D. (2009), 'Evidence for ecological speciation and its alternative', *Science* **323**(5915), 737–741.
URL: *https://science.sciencemag.org/content/323/5915/737*

Skotte, L., Korneliussen, T. S. & Albrechtsen, A. (2013), 'Estimating Individual Admixture Proportions from Next Generation Sequencing Data', *Genetics* **195**(3), 693–702.
URL: *https://doi.org/10.1534/genetics.113.154138*

Smadja, C. M. & Butlin, R. K. (2011), 'A framework for comparing processes of speciation in the presence of gene flow', *MOLECULAR ECOLOGY* **20**(24), 5123–5140.
URL: *https://doi.org/10.1111/j.1365-294X.2011.05350.x*

Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H. & Palumbi, S. R. (2019), 'Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing', *Science* **365**(6452), 487–490.
URL: *https://science.sciencemag.org/content/365/6452/487*

Wellenreuther, M. & Bernatchez, L. (2018), 'Eco-evolutionary genomics of chromosomal inversions', *Trends in Ecology & Evolution* **33**(6), 427–440.
URL: *https://www.sciencedirect.com/science/article/pii/S0169534718300788*

Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M. & Butlin, R. K. (2014), 'Do the same genes underlie parallel phenotypic divergence in different littorina saxatilis populations?', *Molecular Ecology* **23**(18), 4603–4616.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12883*

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., Ravinet, M., Blomberg, A., Mehlig, B., Johannesson, K. & Butlin, R. (2018), 'Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow', *Evolution Letters*

2(4), 297–309.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.74*

Wilder, A. P., Palumbi, S. R., Conover, D. O. & Therkildsen, N. O. (2020), 'Footprints of local adaptation span hundreds of linked genes in the atlantic silverside genome', *Evolution Letters* **4**(5), 430–443.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.189*

Yeaman, S. (2013), 'Genomic rearrangements and the evolution of clusters of locally adaptive loci', *Proceedings of the National Academy of Sciences* **110**(19), E1743–E1751.
URL: *https://www.pnas.org/content/110/19/E1743*

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. (2018), 'PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files', *Bioinformatics* **35**(10), 1786–1788.
URL: *https://doi.org/10.1093/bioinformatics/bty875*

Supplementary information

A Data collection and processing

A.1 Sampling

Littorina saxatilis snails were sampled in March 2017 on the west coast of Galicia in Spain: Centinela (hereafter ER_EA1; N 42° 4' 38.06", W 8° 53' 47.47"). The sample consisted of approximately 600 snails from two 5 m wide bands and separated by 0-5 m, stretching from the upper limit of the species distribution in the splash zone to its lower limit close to low water of spring tides. Sampling was denser in the lower part of the shore where hybridization between 2 ecotypes described before was expected (Galindo et al. 2013). Snails were sampled by hand, haphazardly, as far as possible without reference to phenotype but aiming to avoid juveniles. All of snails were stored in individual tube, moistened with seawater and kept at 4 °C until phenotyping was complete, then the head and foot of each snail was preserved in 100% ethanol.

A.2 DNA sequencing and processing

Low coverage whole genome sequencing (lcWGS) was conducted for 73 snails from ER_EA1, chosen to cover the full sampling range. DNA was extracted using the modified CTAB protocol described by Panova (Panova et al. 2016). Library preparation (in-house high-throughput gDNA library prep) and sequencing (Illumina HiSeq4000, 150bp PE) were carried out by The Oxford Genomics Centre with target coverage 3x based on the estimated genome size of 1.35Gb. Samples were sequenced in eight lanes, for a total of 16 lanes for each individual. Raw reads were trimmed with Trimmomatic v.0.38 (https://github.com/usadellab/Trimmomatic) to remove the Illumina adapters, mapped to the *L. saxatilis* draft reference genome (Swedend, crab; 1.35Gb) (Westram et al. 2018) using bwa mem v.0.7.17 (https://github.com/lh3/bwa) and default settings, and filtered by base and map quality 20 with Samtools v.1.7 (https://github.com/samtools/samtools). PCR duplicates and PE overlap were removed with Picard v.2.0.1 (https://github.com/broadinstitute/picard) and bamUtil v.1.0.15 (http://genome.sph.umich.edu/wiki/BamUtil), respectively. Bam files belonging to the same individual were sorted and merged with Samtools.

B Table and plots

B.1 Mean summary statistics

Linkage group (LG)	LD cluster	Cluster size (cM)	Start (cM)	End (cM)	Mean FST	Mean PBS	Mean Pi	Mean LD (r2)	candidate
LG1	LGC1.1	2.1	0	2.1	0.0019398	0.012677	0.00648093	0.1939411	1
LG1	LGC1.2	5.42	75.53	80.95	0.00219029	0.02014028	0.00536197	0.2134792	1
LG1	LG1	80.95	0	80.95	0.00212763	0.01073747	0.00594011	0.1704242	0
LG2	LGC2.1	13.87	0.34	14.21	0.00614239	0.01879701	0.00576577	0.1999839	1
LG2	LG2	88.76	0	88.76	0.00274161	0.01610663	0.00592564	0.1646505	0
LG3	LG3	68.02	0	68.02	0.00288605	0.01137463	0.00603668	0.1761798	0
LG4	LGC4.1	0.48	1.03	1.51	0.00463436	0.03770791	0.00587387	0.1847306	1
LG4	LG4	56.52	0	56.52	0.00243	0.0140493	0.00587649	0.1711099	0
LG5	LG5	54.07	0	54.07	0.00280035	0.01904912	0.00591619	0.1699921	0
LG6	LGC6.1	29.3	0	29.3	0.00297457	0.03524753	0.00567617	0.1820881	1
LG6	LGC6.2	20.57	8.73	29.3	0.00269045	0.02541398	0.00579544	0.1825618	1
LG6	LG6	60.25	0	60.25	0.00227857	0.01616375	0.00634677	0.1691389	0
LG7	LGC7.1	1.73	36.01	37.74	0.00275902	0.01656839	0.00683345	0.2117835	1
LG7	LGC7.2	9.29	42.08	51.37	0.00364395	0.02627485	0.00641327	0.1799024	1
LG7	LG7	51.37	0	51.37	0.00235047	0.0124258	0.00629948	0.1714745	0
LG8	LG8	54.38	0	54.38	0.0025284	0.01993345	0.00671042	0.186891	0
LG9	LGC9.1	23.18	18.64	41.82	0.00275828	0.02324772	0.00566658	0.1917811	1
LG9	LG9	56.67	0	56.67	0.00256584	0.02090745	0.00763238	0.1833917	0
LG10	LGC10.1	2.54	0.58	3.12	0.00251415	0.01067118	0.00638226	0.1959969	1
LG10	LG10	45.53	0	45.53	0.00280773	0.02139548	0.00636286	0.188335	0
LG11	LGC11.1	0.59	52.32	52.91	0.00259493	0.0145582	0.00686731	0.2100456	1
LG11	LG11	58.39	0	58.39	0.0024994	0.02136298	0.00675471	0.1950033	0
LG12	LGC12.1	26.31	3.32	29.63	0.00155347	0.01451515	0.00667943	0.1775293	1
LG12	LGC12.2	11.52	48.71	60.24	0.00299152	0.01851379	0.00654778	0.1870268	1
LG12	LG12	60.24	0	60.24	0.00314249	0.02937913	0.00609529	0.1726986	0
LG13	LG13	51.3	0	51.3	0.0018803	0.01141404	0.00618406	0.1804347	0
LG14	LGC14.1	11.32	0.39	11.71	0.00325176	0.00927283	0.0060404	0.1966089	1
LG14	LGC14.2	2.9	8.81	11.71	0.00286758	0.00959897	0.00606569	0.1960541	1
LG14	LGC14.3	23.23	11.71	34.94	0.0034732	0.01206736	0.00528926	0.1782203	1
LG14	LG14	48.71	0	48.71	0.00332075	0.00996925	0.00583716	0.1881616	0
LG15	LG15	56.49	0	56.49	0.00201495	0.00941298	0.00633721	0.1755119	0
LG16	LG16	57.44	0	57.44	0.00212438	0.01346309	0.00695419	0.1911745	0
LG17	LGC17.1	15.33	46.99	62.32	0.00561686	0.0379331	0.00655994	0.1748176	1
LG17	LG17	62.32	0	62.32	0.00219622	0.01432516	0.00598865	0.1672243	0

Table 1: Mean summary statistics in inversion candidates and whole chromosome

B.2 PBS

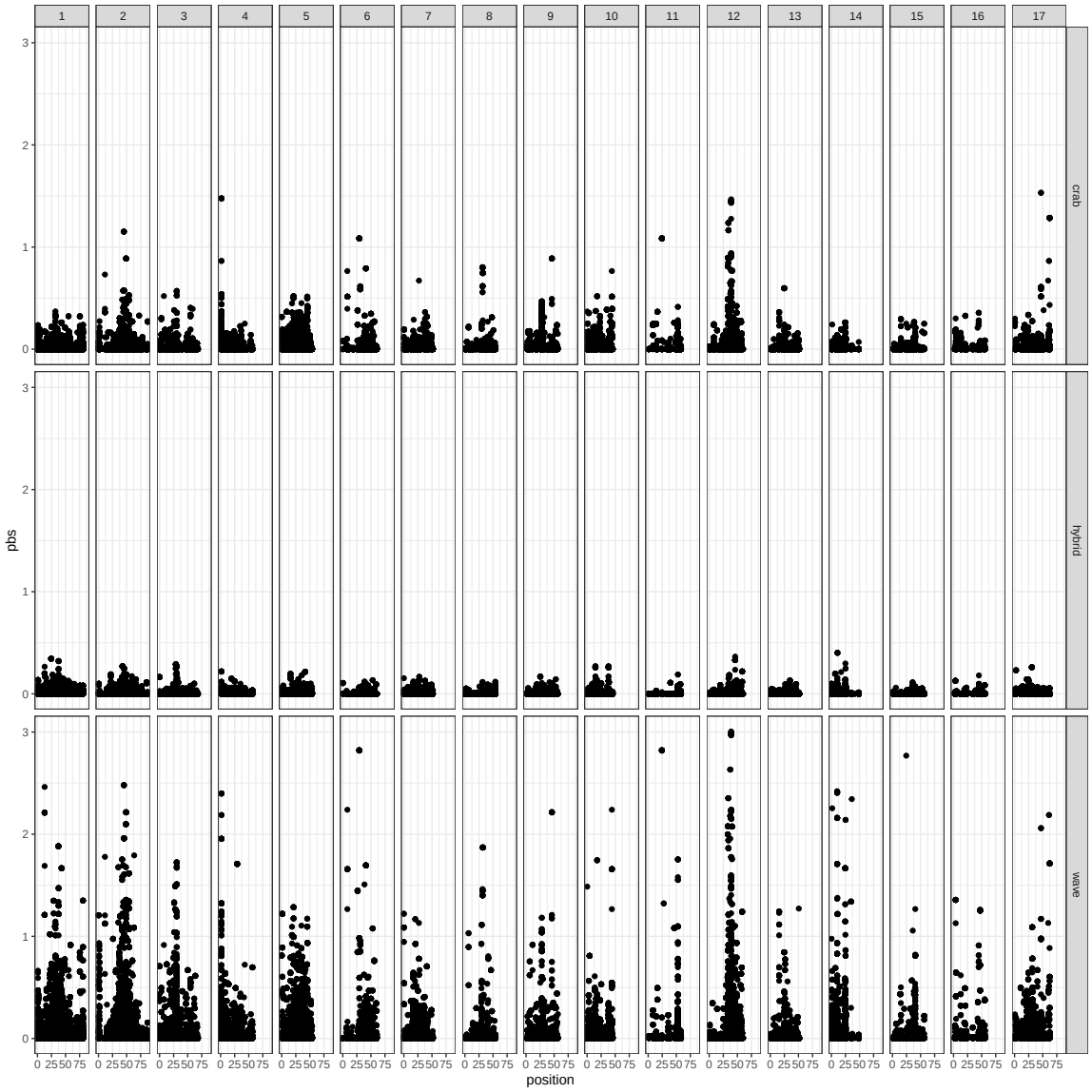

Figure 1: **PBS value scatter plot of every chromosome and population in windows**

B.3 Pi

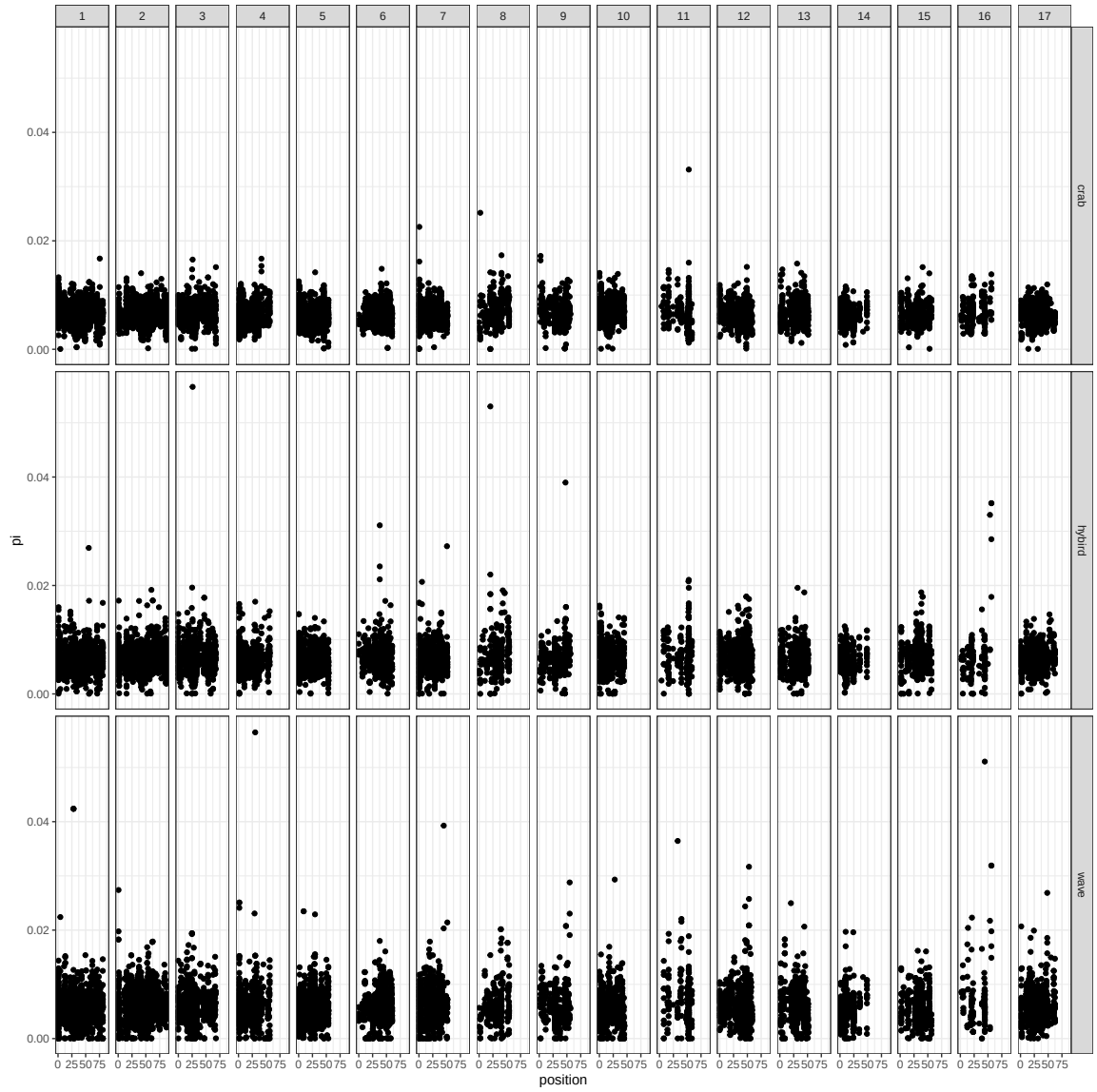

Figure 2: **Pi value scatter plot of every chromosome and population in windows**

B.4 LD decay

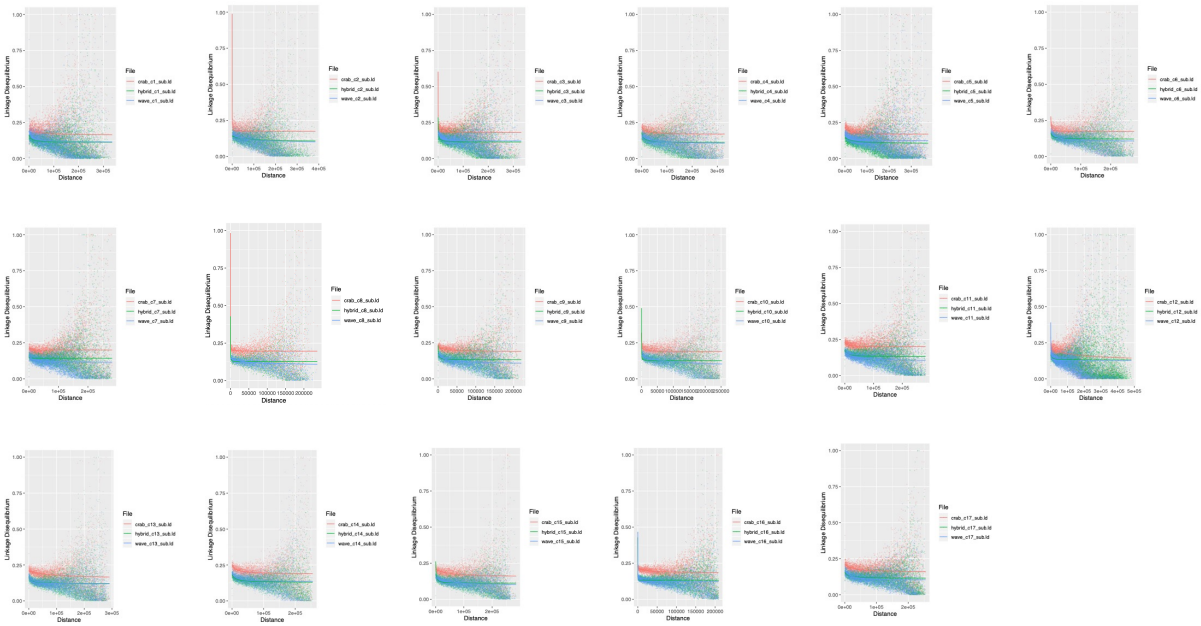

Figure 3: **LD decay analysis of every chromosome**

B.5 LD heat map

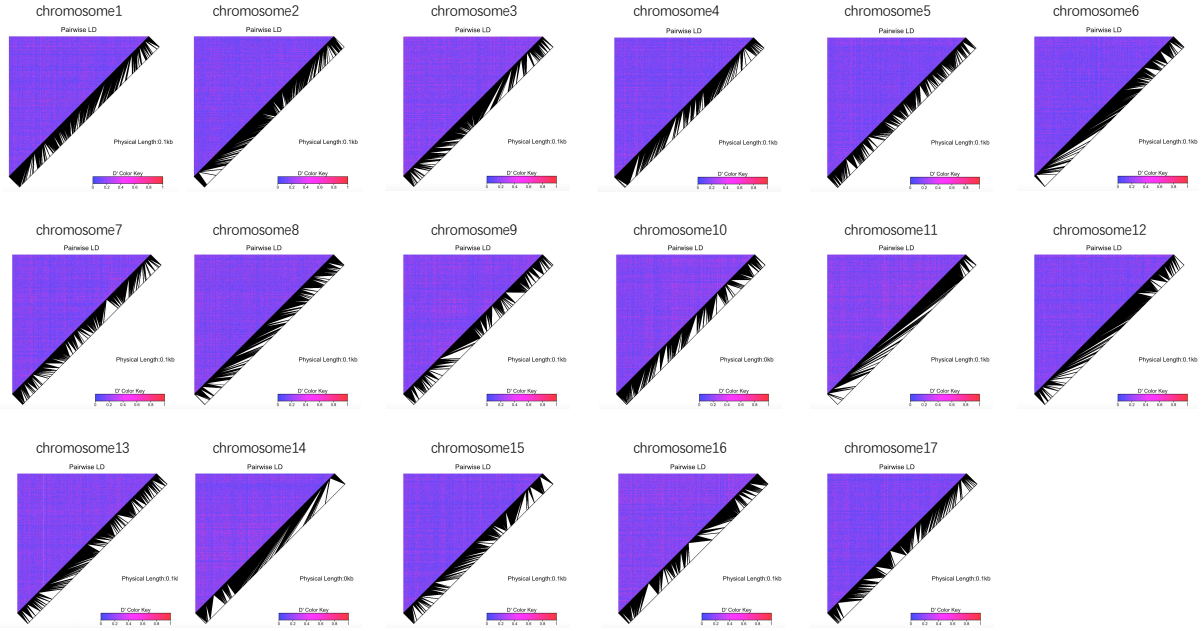

Figure 4: **LD heat map of every chromosome**